Review based Feature Matrix for Predicting ratings in Recommender System

Prafulla Bafna[#1], Shailaja Shirwaikar [*2], Dhanya Pramod [#3] Anagha Vaidya[#4]
[#] Symbiosis International University, Department Of Computer Studies, Pune, India
prafulla,bafna@sicsr.ac.in
Savitribai Phule Pune University, Department of computer studies, Pune, India
scshirwaikar@gmail.com
Symbiosis International University, Department of computer studies, Pune, India
director@scit.edu

***Abstract*—Recommender systems are acquiring extensive popularity and have become essential component of on-line business handling tools because of their capability of providing personalized guidance in selecting products and services. Collaborative Filtering that brings in the popularity aspect of the item amongst the user base, heavily depends on the ratings provided by the users, while Content - based Filtering that brings out the item's features matching the user's taste, requires content information as also user's preference information. Service providers usually invite users to share their experience about the use of service in the form of reviews and ratings. Reviews which are verbose contain a rich source of information about the service's features as also user's preferences while ratings are usually sparse due to user's reluctance to quantify. A feature matrix generated by processing review information using semantic similarity based on synsets can be used along with sparse ratings to generate the complete predicted ratings matrix. The paper presents a modified matrix factorization approach for recommendations using the review based feature matrix**

Keyword- Recommender system, synsets, predictive model.

I. INTRODUCTION

Recommender systems(RS) have become common place in the digital market arena, by providing the on line user of products and services, a personalized selection guidelines. Generally technology used by RS is categorized into two groups as Content- based and Collaborative [1]. Content based systems are based on user profile and item description, giving importance to various features of items that may interest the user or match his taste.It recommends the item that is having the features that user has liked in the previously consumed items [2].Collaborative Filtering harnesses on popularity element or standing of the item across theuser base and works by collecting user feedback in the form of ratings of items. The challenges faced by collaborative filtering are

i. The sparsity of rating matrix due to reluctance on the part of user to rate all the items.

ii. The unavailability of data for new user or new item which is also known as cold start problem.

iii. Hijacking of the RS by pushing own product rating to the higher value and lowering the competitor's rating leading to reduction in the quality of recommendations.

Matrix factorization is widely used, model based approach that estimates ratings. It is a class of latent factor models, where user and items are represented as unknown feature vectors along latent dimensions. The feature vectors are learned using known ratings and learnt features are then used to predict unknown ratings.

In the presence of sparse ratings, reviews in textual form present a rich source of feature information. The reviews are converted into the Feature Matrix with group of terms based on synsets as a feature, utilizing Semantic relationship existing between the terms in the form of Synonyms and meronyms. In this paper Matrix factorization approach uses the above generated feature matrix and the sparse ratings to build the vector model.

The paper is organized as follows. Next section presents the background and work related to Recommender systems. Section 3 presents the processing of reviews using semantic approach to generate feature matrix. Predictive approach is presented in section 4 using matrix factorization model. Section 5 presents the experimental results followed by conclusion and future directions.

II. BACKGROUND

Recommender systems (RS) facilitate prospective buyers to select any product or service having features matching his or her choice as also giving consideration to the popularity amongst other users [3][4]. It has varied applications like in purchasing product (Amazon), listening a song (Last.fm) or selecting a hotel (TripAdvisor) etc. [2].

RS uses three common techniques Collaborative Filtering(CF), Content based filtering and more recently a combination of above two techniques that is Hybrid filtering. Collaborative Filtering uses two approaches neighborhood based and model based. Neighborhood based methods are user based or item based. The user-based methods use ratings to link a user with a set of like- minded users. It recommends to the new user a set of items that are liked by her/his neighbors; in item-based method, items that are similar to those that a user has viewed/purchased before, are recommended.

Model-based CF focuses on learning the latent factors that represent users' inherent preferences over an item's multiple dimensions [5]. Model based methods perform well when there is sufficient rating information.

Most product and service providers collect the user feedback in qualitative form as a textual review or quantified form as ratings. While some users find it easy to rate a service, most users prefer to share their experience of the use of service as a review. While ratings can be easily manipulated, reviews represent user sentiments in a more reliable manner [6]. Recently several researchers have used the valuable information hidden in reviews to address the rating problems. The most commonly used approach is to identify frequently occurring terms in reviews as indicators of item as well as reviewer characteristics [7]. Another approach is to identify review topics which can be carried out by using a frequency-based approach on extracted terms or phrases [8] or using topic modeling approach such as Latent Dirichlet Allocation(LDA) [9]. Opinion mining is another approach where positive or negative sentiments of the user about the item can be identified by aggregating sentiments of all opinion words[10]. The helpful reviews as voted by other users can be given higher weightage thus improving predictions [11].

Matrix factorization methods are used as unsupervised learning methods for latent variable decomposition and dimensionality reduction. It maps both users and items to a joint latent factor space of dimensionality [12]. A vector qi is related with item i and a vector pu is linked to each user u

qi is the degree to which the item has the features.

pu is the degree to which user is interested in item features

The result of dot product qi Tpu is termed as the interaction between user u and item i – the user's overall interest in the item's characteristics rui denotes user u's rating to item i, The matrices can be used to compute the recommendation score for any user and item [13][14][15].Machine learning techniques can be used to generate the parameters that govern the relationship between item and features as in qi and also between user and features that is pu using known ratings data. In this paper, the review based feature matrix can replace qi while the available sparse ratings can be used to generate the parameters that govern the relationship between the users and item features.

III.REVIEW BASED FEATURE MATRIX GENERATION

The first step in review based recommendation is to generate the feature matrix. The reviews are preprocessed by removing stop words, stemming list etc[18][19]. All the terms with their frequencies are identified and term set is generated. Synonyms and Meronyms are identified for each term and assembled into groups to form the synsets. The synsets having group frequency greater than the threshold frequency are chosen to represent the columns and reviews are placed in rows [20]. The synset document matrix containing the group frequency count is normalized to form the feature matrix.

A. Data Collection

The dataset used is from the tar file having 12,773 reviews of hotels. These were downloaded from the tripadvisor site. The data is in JSON format. [http://sifaka.cs.uiuc.edu/~wang296/Data/index.html].

In each hotel's data file, there are approximately 100-300 reviews of that particular hotel by different users. The dataset contains ratings and reviews, author name, location including country and state, short review1, short review2, author review ID, Name of hotel, Descriptive review. The Hotel table contains id, name, URL, price, address etc. The problem related to dataset is that the reviews are quite extensive and the ratings are very sparse.

B. Feature Matrix generation

Feature Matrix (X) is a number of hotels (nh) by number of features (nf) matrix containing the normalized feature quantification for each hotel and each set of features.Reviews contains terms that describe the features of an item or service. Though a single feature indicator term, may not be frequent enough in a review but may be repetitively used across the reviews. The different users may not use the same terms but may use synonym

or meronym of the same term. The synset grouping thus helps in bringing together different terms indicating the same feature. As reviews contained a lot of location specific and other hierarchies, both synonyms and meronyms were effectively used in forming the synset groups.

Table 1 shows the feature matrix containing normalized values for 28 features and 10 hotels. Table 2 shows the feature matrix containing normalized values for 34 features and 10 hotels.

For comparative study the feature matrix is generated using two approaches, in the first approach ten reviews per hotel were considered and in the second approach 40 reviews were considered for each hotel. The feature matrices generated by the first and second approach are shown in Table I and Table II respectively. In the second approach synset grouping becomes stronger because synonyms of terms are repeated across reviews

of same hotel thus strengthening the frequency count. The number of features increased from 22 for single review to 28 for 10 reviews and 34 for 40 reviews as shown in various shading in Table VI

TABLE I Feature Matrix using ten reviews per hotel

	F1	**F2**	**F3**	**F4**	**…**	**F27**	**F28**
H1	0.455	0.273	0.273	0.273	..	0.091	0.818
H2	0.455	0.182	0.273	0.364	..	0.091	0.727
H3	0.091	0.636	0.182	0	..	0.364	0.091
H4	0.364	0.273	0.273	0.273	..	0.091	0.636
H5	0	0.091	0.091	0.182	..	0.818	0.273
H6	0.364	0.182	0.182	0.273	..	0.091	0.455
H7	0	0.091	0.091	0.182	..	0.818	0.273
H8	0.091	0.636	0.182	0	..	0.273	0.091
H9	0	0.091	0.091	0.182	..	0.818	0.273
H10	0.091	0.636	0.182	0	..	0.182	0.182

TABLE II FEATURE MATRIX USING FORTY REVIEWS OF ONE HOTEL

	F1	**F2**	**F3**	**F4**	**..**	**F33**	**F34**
H1	0.231	0.692	0	0.923	..	0.07	0.39
H2	0.154	0.308	0.385	0.308	..	0.07	0.28
H3	0.231	0.769	1	0.538	..	0.28	0
H4	0.308	0.692	0	0.923	..	0.07	0.21
H5	0.385	0.154	0.077	0.231	..	0.629	0.14
H6	0.462	0.308	0	0.308	..	0.07	0.38
H7	0.538	0.154	0.077	0.231	..	0.629	0.12
H8	0.615	0.769	1	0	..	0.21	0
H9	0.692	0.154	0.077	0.231	..	0.629	0.14
H10	0.769	1	0.846	0	..	0.14	0

TABLE III RATING MATRIX

	H1	**H2**	**H3**	**H4**	**H5**	**H6**	**H7**	**H8**	**H9**	**H10**
u1	?	?	?	?	2	2	?	?	?	?
u2	3	?	4	?	?	?	?	?	?	?
u3	?	?	?	?	?	2	?	?	?	3
u4	?	?	?	?	?	1	?	?	?	5
u5	?	?	?	?	4	?	?	?	5	?
u6	?	?	?	?	3	2	?	?	?	?
u7	?	?	?	?	?	?	?	3	2	?
u8	?	?	?	?	5	5	?	?	?	?
u9	?	?	?	?	?	?	?	?	4	1
u10	?	4	3	4	3	2	1	3	4	1

Fig. 1. A sample line graph using colors which contrast well both on screen and on a black-and-white hardcopy

IV. PREDICTIVE MODELING APPROACH

In this section the predictive approach is presented for recommendations using the feature matrix and the available user ratings. For each user we need to learn the parameter vector Theta that governs the relationship between the features and the rating.

Rating (Y) is a number of users (nu) by number of hotels (nh) matrix containing the rating given by user to the hotel. The rating is usually a number between 1 and 5. There may be several missing values which are indicated by a '?'. The table III shows the sparse rating matrix.

For computational purpose an Indicator matrix (R) is used which is a binary valued matrix used to indicate the presence and absence of ratings. If the user i has rated hotel j, thenR(i, j)=1 and 0 otherwise.

Parameter Matrix (θ) is a number of user (nu) by number of feature (nf) matrix containing the parameter values that govern the relationship between the features and ratings.

Predicted Ratings (P) is a number of users (nu) by number of hotels (nh) matrix containing the predicted ratings for each user and each hotel. The P matrix can be computed once the parameter values are available using the following formula.

P = θXTranspose

For learning the parameters the Gradient descent approach is used

A. *Gradient Descent Approach*

The machine learning algorithm tries to minimize the cost function. The cost is computed as the mean squared error between the predicted and the actual ratings [16][17].

Mean squared error (J) can be computed using Equation (1)

$$J = \frac{1}{2}\sum_{(i,j):R(i,j)=1}(\theta X^{\text{Transpose}} - y)^2$$

The gradient decent approach iteratively modifies the θ parameters by adding the gradient which is the partial derivative of the error.

Gradient (grad) for each parameter θ can be computed using the Equation (2)

$$Grad_k^j = \sum_{(i,j):R(i,j)=1}((\theta^j)^T X^i - y^{(i,j)})X_k^i \qquad (2)$$

At each iteration the error decreases, so that after finite number of iterations, the parameters that best fit the data can be obtained. The cost function and gradient are modified by using the Equation (3)

Choosing Regularization parameter lambda (λ)

To avoid overfitting of the model to the training data, regularization parameter governed by lambda is

$$J = J + \frac{\lambda}{2}\sum_{j=1}^{n_u} \sum_{k=1}^{n_f} (\theta_k^i)^2$$

$$Grad_k^j = Grad_k^j + \lambda\theta_k^j \qquad (3)$$

The lambda (λ) parameter is chosen after executing algorithm for different lambda (λ) values.[17]

The learning curve for for different values of lambda (λ) for ten reviews based feature matrix is shown in the fig 1. and the learning curve for training and cross validation error for different values of lambda (λ).

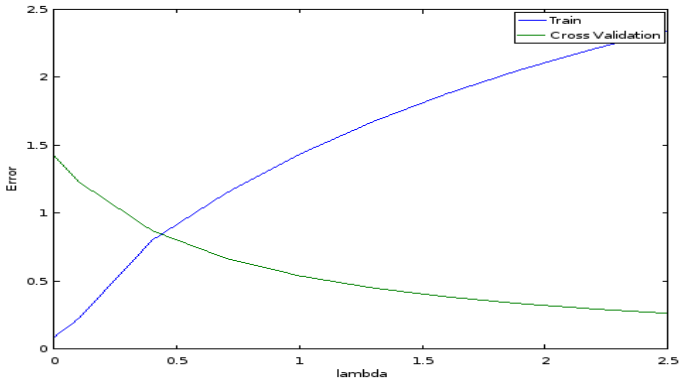
B. *Predicted Ratings*

The matrix factorization algorithm is implemented on octave platform and training is carried out using built-in octave functions.

The gradient decent algorithm was executed using two feature matrices. Table IV and Table V shows predicted ratings for the feature matrix which is based on a ten reviews and forty reviews respectively. Variations in the predicted ratings can be clearly observed in Table V.

TABLE IV PREDICTED RATING MATRIX BASED ON TEN REVIEWS

	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10
U1	2	3.27	2	3.27	2	2	0.55	2.4	2	1.91
U2	3	4.24	3	4.24	3	3	1.59	3.4	3	2.94
U3	2	3.28	2	3.28	2	2	0.53	2.4	2	1.56
U4	5	6.26	5	6.26	5	5	3.57	5.4	5	4.92
U5	4	4.59	4	4.59	4	4	3.33	4	4	3.91
U6	2	3.2	2	3.2	2	2	0.63	2.3	2	1.94
U7	2	2.57	2	2.57	2	2	1.35	2	2	1.92
U8	5	6.22	5	6.22	5	5	3.61	5.3	5	4.24
U9	4	4.59	4	4.59	4	4	3.33	4	4	3.92
U10	2.6	3.82	3	3.82	2.6	2.6	1.2	2.9	2.6	1.71



learning curve: Ten reviews per hotel based feature matrix

TABLE V PREDICTED RATING MATRIX BASED ON FORTY REVIEW

	H1	H2	H3	H4	H5	H6	H7	H8	H9	H10
U1	2	3.312	2	3.312	2	2	0.5	2.366	2	1.92
U2	3	4.278	3	4.278	3	3	1.54	3.368	3	2.95
U3	2	3.327	2	3.327	2	2	0.48	2.366	2	1.59
U4	5	6.297	5	6.297	5	5	3.52	5.373	5	4.94
U5	4	4.606	4	4.606	4	4	3.31	4.03	4	3.92
U6	2	3.24	2	3.24	2	2	0.58	2.344	2	1.96
U7	2	2.591	2	2.591	2	2	1.32	2.024	2	1.94
U8	5	6.259	5	6.259	5	5	3.56	5.339	5	4.19
U9	4	4.606	4	4.606	4	4	3.31	4.028	4	3.93
U10	2.6	3.862	2.6	3.862	2.6	2.6	1.16	2.951	2.6	1.63

TABLE VI FEATURE MATRIX

Sr.no	Features
1	good, clean
2	nice, fantastic
3	breakfast, food
5	guest, visitor ,customer
6	romance, love
..	..
21	Italy, Rome
22	Australia, Victoria, Tasmania
23	California, Berkeley
24	recommend, suggest
25	home, house
26	care , concern
27	microwave
28	London , Whitehall
29	airport
30	distance
31	facility
32	close near
33	Canada, Manitoba, Nunavut
34	Africa, Barbary

V. CONCLUSION

The paper presents an approach to predict the ratings based on reviews when ratings are sparse. Semantic similarity between the terms is used to generate the feature matrix from reviews. The predicted rating matrix produced using two types of feature matrices. The variations in the predicted ratings are presented. The approach is based on modified matrix factorization for recommendations using the feature

REFERENCES

[1] Sarwar, Badrul, George Karypis, Joseph Konstan, and John Riedl. "Item-based collaborative filtering recommendation algorithms." In Proceedings of the 10th international conference on World Wide Web, pp. 285-295. ACM, 2001
[2] Chen, Li, Guanliang Chen, and Feng Wang. "Recommender systems based on user reviews: the state of the art." User Modeling and User-Adapted Interaction 25, no. 2 (2015): 99-154.
[3] Dong, Ruihai, Michael P. O'Mahony, Markus Schaal, Kevin McCarthy, and Barry Smyth. "Sentimental product recommendation." In Proceedings of the 7th ACM conference on Recommender systems, pp. 411-414. ACM, 2013.
[4] Esparza, Sandra Garcia, Michael P. O'Mahony, and Barry Smyth. "Effective product recommendation using the real- time web." In Research and Development in Intelligent Systems XXVII, pp. 5-18. Springer London, 2011.
[5] Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems." Computer 42, no. 8 (2009): 30-37
[6] Musat, C-C., Yizhong Liang, and Boi Faltings. "Recommendation using textual opinions." In IJCAI International Joint Conference on Artificial Intelligence, no. EPFL-CONF-197487, pp. 2684-2690. 2013.
[7] Manning CD, Raghavan P, Sch➧utze H (2008) Introduction to Information Retrieval. Cambridge University Press,1st Edition
[8] McAuley, Julian, and Jure Leskovec. "Hidden factors and hidden topics: understanding rating dimensions with review text." In Proceedings of the 7th ACM conference on Recommender systems, pp. 165-172. ACM, 2013
[9] Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of machine Learning research 3, no. Jan (2003): 993-1022
[10] Wang, Yuanhong, Yang Liu, and Xiaohui Yu. "Collaborative filtering with aspect-based opinion mining: A tensor factorization approach." In 2012 IEEE 12th International Conference on Data Mining, pp. 1152-1157. IEEE, 2012.
[11] Raghavan, Sindhu, Suriya Gunasekar, and Joydeep Ghosh. "Review quality aware collaborative filtering." In Proceedings of the sixth ACM conference on Recommender systems, pp. 123-130. ACM, 2012.
[12] Bokde, Dheeraj, Sheetal Girase, and Debajyoti Mukhopadhyay. "Matrix factorization model in collaborative filtering algorithms: A survey." Procedia Computer Science 49 (2015): 136-146.
[13] Jamali, Mohsen, and Martin Ester. "A transitivity aware matrix factorization model for recommendation in social networks." In IJCAI, vol. 11, pp. 2644-2649. 2011.

[14] Garcia Esparza, Sandra, Michael P. O'Mahony, and Barry Smyth. "A multi-criteria evaluation of a user generated content based recommender system." In Presented at the 3rd Workshop on Recommender Systems and the Social Web (RSWEB-11), 5th ACM Conference on Recommender Systems, Chicago, IL, USA, 23-27 October 2011. 2011.
[15] Hariri, Negar, Yong Zheng, Bamshad Mobasher, and Robin Burke. "Context-aware recommendation based on review mining." General Co-Chairs (2011): 27.
[16] Huang, Guang-Bin, Qin-Yu Zhu, and Chee-Kheong Siew. "Extreme learning machine: theory and applications." Neurocomputing 70, no. 1 (2006): 489-501.
[17] http://www.holehouse.org/mlclass/16_Recommender_Syste ms.html: Retrieved November 2015
[18] Jannach, Dietmar, Markus Zanker, Alexander Felfernig, and Gerhard Friedrich. Recommender systems: an introduction. Cambridge University Press, 2010.
[19] Wang, Yuanhong, Yang Liu, and Xiaohui Yu. "Collaborative filtering with aspect-based opinion mining: A tensor factorization approach." In 2012 IEEE 12th International Conference on Data Mining, pp. 1152-1157. IEEE, 2012.
[20] Bafna, Prafulla Bharat, Shailaja Shirwaikar, and Dhanya Pramod. "Multi-Step Iterative Algorithm for Feature Selection on Dynamic Documents." International Journal of Information Retrieval Research (IJIRR) 6, no. 2 (2016): 24-40